Application transmittal

A

UTILITY PATENT APPLICATION TRANSMITTAL

(only for new non-provisional applications under 37 CFR 1 53(b))

Attorney Docket No. **Beutnagel 4-1-13-3**

First Named Inventor or Application Identifier **Mark Beutnagel**

Title **Integration of Talking Heads and Text-to-Speech Synthesizers for Visual TTS**

Express Mail Label no. **EM164542404US**

JC511 U.S. PTO 09/224583 12/31/98

12/31/98 JC511 09/224583

To: **Assistant Commissioner for Patents**
Box Patent Application
Washington D.C. 20231

APPLICATION ELEMENTS	ACCOMPANYING APPLICATION PARTS
☒ Fee Transmittal Form (original and duplicate)	☐ Assignment
☒ Specification Total Pages **19** title cross reference to related applications (e.g. provisional application) background summary brief description of the drawings (if filed) detailed description claims abstract	☐ Recordation form ☐ Power of Attorney ☒ Postcard ☐ Small entity statement ☐ Certified copy of priority documents ☐ Information disclosure statement ☐ Copies of IDS citations ☐ 37 CFR 3.73(b) Statement ☒ Check ☐ Other
☒ Drawing(s) Total Pages **1**	
☒ Declaration Total Pages **3**	
a. ☐ Newly executed	
b. ☐ Copy from a prior application (37 CFR 1.63(d)) (for continuations/divisionals with section below filled out)	
☐ Deletion of Inventor(s) Signed Statement attached deleting inventor(s) named in the prior application. 37 CFR 163 (d)(2) and 1 33(b).	
☐ Incorporation by reference (usable if Declaration is a copy): The entire disclosure of the prior application, from which a copy of the oath or declaration is supplied, is considered as being part of the disclosure of the accompanying application is hereby incorporated by reference herein.	
☐ Other	

If a CONTINUING APPLICATION, check appropriate box and supply the requisite information:

☐ Continuation ☐ Divisional ☐ Continuation-in-part (CIP) of prior Application No:

CORRESPONDENCE ADDRESS

☐ Customer Number or Bar Code Label	(insert Customer No. or Attach bar code label here)	☒ Correspondence Address below

NAME	**Samuel H. Dworetsky**
ADDRESS	**AT&T Corp. P.O. Box 636** **Middletown, NJ 07748-4801**
COUNTRY	**United States** FAX **(732) 957-5505**

SIGNATURE OF APPLICANT ATTORNEY, OR AGENT

Name	**Henry T. Brendzel**	Reg. No.	**26,844**
Telephone	**(973) 467-2025**		
Signature	Henry Brendzel	Date	12/31/98

I hereby certify that this Application is being deposited with the United States Postal Service "Express Mail Post Office to Addressee" service under 37 CFR 1.10 on the date indicated above and is addressed to the Assistant Commissioner for Patents, Washington D.C. 20231.

12/31/98	**Henry T. Brendzel**	Henry Brendzel
Date of Deposit	(Printed Name of Person Mailing Paper)	(Signature of Person Mailing Paper)

Integration of Talking Heads and Text-to-Speech Synthesizers for Visual TTS

Reference to a Related Application

This invention claims the benefit of provisional application No. 60/082,393, filed April 20, 1998, titled "FAP Definition Syntax for TTS Input."

Background of the Invention

The success of the MPEG-1 and MPEG-2 coding standards was driven by the fact that they allow digital audiovisual services with high quality and compression efficiency. However, the scope of these two standards is restricted to the ability of representing audiovisual information similar to analog systems where the video is limited to a sequence of rectangular frames. MPEG-4 (ISO/IEC JTC1/SC29/WG11) is the first international standard designed for true multimedia communication, and its goal is to provide a new kind of standardization that will support the evolution of information technology.

MPEG-4 provides for a unified audiovisual representation framework. In this representation, a scene is described as a composition of arbitrarilyy shaped audiovisual objects (AVOs). These AVOs can be organized in a hierarchical fashion, and in addition to providing support for coding individual objects, MPEG-4 also provides facilities to compose that hierarchical structure.

One of these AVOs is the Face Object, which allows animation of synthetic faces, sometimes called Talking Heads. It consists of a 3D synthetic visual object representing a human face, a synthetic audio object, and some additional information required for the animation of the face. Such a scene can be defined using the BInary Format for Scene (BIFS), which is a language that allows composition of 2D and 3D objects, as well as animation of the objects and their properties.

The face model is defined by BIFS through the use of nodes. The Face Animation Parameter node (FAP) defines the part of the face has to be animated. The Face Description Parameter node (FDP) defines the rules to animate the face model. The audio object can be natural audio, or created at the decoder with some proprietary Text-To-Speech (TTS). In the case of an encoded stream containing natural audio, an independent

FAP stream drives the animation, and time stamps included in the streams enable the synchronization between the audio and the animation.

A TTS is a system that accepts text as input, and outputs an intermediate signal that comprises phonemes, and the final signal that comprises audio samples corresponding to the text. MPEG-4 does not standardize the TTS Synthesizer, but it provides a Text-To-Speech Interface (TTSI). By sending text to the decoder, the animation is driven by the FAP stream and by the TTS.

MPEG-4 defines a set of 68 Face Animation Parameters (FAPs), each corresponding to a particular facial action that deforms a face from its neutral state. These FAPs are based on the study of minimal perceptible actions, and are closely related to muscle action. The value for a particular FAP indicates the magnitude of the corresponding action. The 68 parameters are categorized into 10 groups, as shown in Table 1 of the appendix. Other than the first group, all groups are related to different parts of the face. The first group contains two high-level parameters (FAP 1 and FAP 2); visemes and expressions. A viseme is a visual version of a phoneme. It describes the visually distinguishable speech posture involving the lips, teeth and tongue. Different phonemes are pronounced with a very similar posture of the mouth, like "p" and "b" and, therefore, a single viseme can be related to more than one phoneme. Table 2 in the appendix shows the relation between visemes and their corresponding phonemes.

In order to allow the visualization of mouth movement produced by coarticulation, transitions from one viseme to the next are defined by blending the two visemes with a weighting factor that changes with time along some selected trajectory.

The expression parameter (FAP 2) defines 6 high level facial expressions, such as joy, sadness, anger, etc. They are described in Table 3 of the appendix. The nine other FAP groups, which represent FAP 3 to FAP 68, are low-level parameters, like move left mouth corner up.

Each FAP (except FAP1 and FAP2) is defined in a unit, which can vary from one parameter to another. Unlike visemes and expressions, each low-level FAP characterizes only a single action. Therefore, a low-level action is completely defined with only two numbers, the FAP number, and the amplitude to be applied to the action. In the case of

high-level parameters, a third number, called FAPselect, is required to determine which viseme (in case of FAP 1), or which expression (in case of FAP 2) is to be applied.

For each frame, the receiver applies and performs the deformations on the face model using all FAPs. Once all actions have been done on the model, the face is rendered.

MPEG-4 allows the receiver to use a proprietary face model with its own animation rules. Thus, the encoder sends signals to control the animation of the face by sending FAPs but has no knowledge concerning the size and proportion of the head to animate, or any other characteristic of the decoding arrangements. The decoder, for its part, needs to interpret the values of the FAPs in a way such that the FAPs produce reasonable deformation. Because the encoder is not aware of the decoder that will be employed, the MPEG-4 standard contemplates providing *normalized* FAP values in face animation parameter units (FAPU). The FAPU are computed from spatial distances between key facial features on the model in its neutral state, such as iris diameter, eye separation, eye-to-nose separation, Mouth-to-nose separation, and Mouth width.

FIG. 1 presents a block diagram of a prior art face rendering arrangement that employs the FAP information that is available with MPEG-4. It includes an audio signal on line 10 that is applied to decoder 100 and thence to synthesizer 120, and a FAP stream on line 11 that is applied to face rendering module (FRM) 110. Module 110 can be a separate piece of hardware, but often it is a software module that is executed on a processor. A face model and its animation rules may be applied to FRM 110 via line 12. While decoder 100 decodes the audio signal and synthesizer 120 synthesizes it, FRM 110 concurrently renders the face based on the applied FAP stream. Compositor 130, responsive to synthesizer 120 and FRM 110, simultaneously plays the audio and the animated model video that result from applying the FAPs to FRM 110. Synchronization is achieved at the decoder by retrieving timing information from the streams. This timing information is of two types, and must be included in the transmitted streams. The first type is used to convey the speed of the encoder clock, while the second one consists of time stamps attached to portions of the encoded data.

Providing for this synchronization (between what is said and the desired facial expressions) on the encoder side is not trivial, and the problem is certainly not reduced when a TTS arrangement is contemplated. The reason lies in the fact that whereas faces

are animated at constant frame rate, the timing behavior of a TTS Synthesizer on *the decoder side* is usually unknown. It is expected that there will be a very large number of commercial applications where it will be desirable to drive the animation from a text. Therefore, solving the synchronization problem is quite important.

Summary of the Invention

An enhanced arrangement for a talking head driven by text is achieved by sending FAP information to a rendering arrangement that allows the rendering arrangement to employ the received FAPs in synchronism with the speech that is synthesized. In accordance with one embodiment, FAPs that correspond to visemes which can be developed from phonemes that are generated by a TTS synthesizer in the rendering arrangement are not included in the sent FAPs, to allow the local generation of such FAPs. In a further enhancement, a process is included in the rendering arrangement for creating a smooth transition from one FAP specification to the next FAP specification. This transition can follow any selected function. In accordance with one embodiment, a separate FAP value is evaluated for each of the rendered video frames.

Brief Description of the Drawing

FIG. 1 depicts a prior art rendering arrangement that is useful for rendering a talking head from an audio stream and a separate FPAs stream;

FIG. 2 presents an arrangement where phonemes developed by the TTS synthesizer of FIG. 1 are employed to develop visemes locally; and

FIG. 3 shows an arrangement where FAP information is embedded in the incoming TTS stream.

Detailed Description

FIG. 1 depicts a prior art rendering arrangement that receives signals from some encoder source and develops therefrom an audio signal and a talking head video. More specifically, the rendering arrangement of FIG. 1 is arranged to be useful for TTS systems as well as for natural audio. The difference between a natural audio system and a TTS system lies in element 100, which converts an incoming text string into speech. When

element 100 is responsive to natural audio, it is effectively a decoder. When elements 100 is responsive to ASCII text, it is effectively a TTS synthesizer.

One enhancement that is possible, when employing the FIG. 1 arrangement to synthesize speech is to use the phoneme information (the phoneme's identity, its start time, and its duration) that is generated as an intermediate output of the TTS synthesizer to generate some viseme FAPs. The generated FAPs are assured to be fairly well synchronized with the synthesized speech and, additionally, the local generation of these FAPs obviates the need to have the encoder generate and send them. This enhanced arrangement is shown in FIG. 2, and it includes a phoneme to FAP converter 140 that is interposed between decoder 100 and FRM 110.

As indicated above, the synchronization between the generated visemes and the speech is fairly good. The only significant variable that is unknown to FRM 110 is the delay suffered between the time the phonemes are available and the time the speech signal is available. By comparison, the synchronization between the incoming FAP stream and the synthesized speech is much more problematic. As indicated above, MPEG-4 does not specify a standard for the operation of TTS equipment, but specifies only a TTS Interface (TTSI). Therefore, the precise characteristics of the TTS synthesizer that may be employed in the FIG. 2 arrangement are not known. The encoder that generates the FAP stream does not know whether a receiving decoder 100 will create speech that is fast, or slow, at a constant rate or at some variable rate, in monotone or is "sing-song," etc. Consequently, synchronization between the FAP stream and the output of the TTS synthesizer is usually not very good.

We have concluded that a better approach for insuring synchronization between the TTS synthesizer 120 and the output of FRM 110 is to communicate prosody and timing information to TTS synthesizer 120 along with the text and in synchronism with it. In our experimental embodiment this is accomplished by sending the necessary FAPs stream (i.e., the entire FAPs stream, minus the viseme FAPs that would be generated locally by converter 140) embedded in the TTS stream. The FAPs information effectively forms bookmarks in the TTS ASCII stream that appears on line 10. The embedding is advantageously arranged so that a receiving end could easily cull out the FAP bookmarks from the incoming streams.

This enhanced arrangement is shown in FIG. 3, which differs from FIG. 2 in that it includes an enhanced decoder, 150. Decoder 150 extracts the FAPs stream contained in the TTS stream on line 10 and applies the extracted FAPs stream to converter 140 via line 13. The function of converter 140 in FIG. 3 is expanded to not only convert phoneme information into FAPs but to also merge the developed FAPs with the FAPs that are extracted by decoder 150 from the incoming TTS stream and provided to converter 140.

Illustratively, the syntax of the FAPs bookmarks is **<FAP # (FAPselect) FAPval FAPdur>**, where the # is a number that specifies the FAP, in accordance with Table 4 in the appendix. When the # is a "1", indicating that it represents a viseme, the **FAPselect** number selects from Table 1. When the # is a "2", indicating that it represents an expression, the number selects from Table 2. **FAPval** specifies the magnitude of the FAP action, and **FAPdur** specifies the duration.

Simply applying a FAP of a constant value and removing it after a certain amount of time does not give a realistic face motion. Smoothly transitioning from one FAP specification to the next FAP specification is much better. Accordingly, it is advantageous to include a transitioning schema in the FIG. 3 arrangement; and in accordance with one such schema, the **FAPval** defines the value of the FAP to be applied at the end of **FAPdur**. The value of the FAP at the beginning of the action (startValue) depends on the previous value and can be equal to:

- 0 if the FAP bookmark sequence is the first one with this FAP #
- **FAPval** of the previously applied FAP, if a time longer than the previous **FAPdur** has elapsed between the two FAP specifications.
- The actual reached value due to the previous FAP specification, if a time shorter than the previous **FAPdur** has elapsed between the two FAP specifications.

To reset the action, a FAP with **FAPval** equal to 0 may be applied.

While having a linear transition trajectory from one FAP to the next is much better than an abrupt change, we realized that any complex trajectory can be effected. This is achieved by specifying a FAP for each frame, and a function that specifies the transition trajectory from the FAP from frame to frame. For example, when synthesizing a phrase such as "...really? You don't say!" it is likely that an expression of surprise will be assigned to, or associated with, the word "really," and perhaps for some time after the next word, or

words are synthesized. Thus, this expression may need to last for two seconds or more, but the FAP that specifies surprise is specified only once by the source.

A trajectory for fading of the previous expression and for establishment of the "surprise" expression needs to be developed for the desired duration, recognizing that the next expression may be specified before the desired duration expires, or some time after the desired duration expires. Thus, the FIG. 3 rendering arrangement needs choose the aforementioned trajectory. In accordance with this invention, any desired trajectory can be established from the starting time throughout the *FAPdur* interval, and beyond. One way to accomplish this is to select a function that is evaluated at every frame to yield strength, or magnitude, of the expression (e.g., big smile, or small smile) at every frame that is rendered. The function can be linear, as described above, but it can also be a non-linear function. Of course, one need not and restrict oneself to use only some selected function. That is, going from expression A to expression B need not follow a function that is the same as the function followed when going from expression B to expression C.

We have identified a number of useful transition trajectory functions. They are:

$$f(t) = a_s + (a - a_s)t\,; \tag{1}$$

$$f(t) = a_s + (1 - e^{-t})(a - a_s)\,, \tag{2}$$

$$f(t) = a_s + \frac{(a - a_s)}{1 - e^{-\lambda(t - \frac{FAPdur}{2})}}\,, \text{ and} \tag{3}$$

$$f(t) = a_s(2t^3 - 3t^2 + 1) + (-2t^3 + 3t^2)a + (t^3 - 2t^2 + t)g_s\,, \tag{4}$$

with $t = [0, 1]$, the amplitude a_s at the beginning of the FAP, at t=0, control parameter λ and the gradient g_s of $f(0)$ with is the FAP amplitude over time at t=0. If the transition time $T \neq 1$, the time axis of the functions need to be scaled, since these functions depend only on a_s, λ, g_s, and T, and thus are completely determined as soon as the FAP bookmark is known.

The most important criterion for selecting a transition trajectory function is the resulting quality of the animation. Experimental results suggest that when linear interpolation is used, and when equation (2) is used, sharp transitions result in the combined transition trajectory, which do not result in a realistic rendering. Equations (3) and (4) yield better results. On balance, we have concluded that the function of equation

(4) gives the best results, in terms of realistic behavior and shape prediction. This function enables one to match the tangent at the beginning of a segment with the tangent at the end of the previous segment, so that a smooth curve can be guaranteed. The computation of this function requires 4 parameters as input, which are: the value of the first point of the curve (*startVal*), its tangent (*startTan*), the value to be reached at the end of the curve (equal to *FAPVal*) and its tangent.

For each FAP #, the first curve (due to FAP # bookmark$_{i=0}$) has a starting value of 0 ($startVal_{i=0} = 0$) and a starting tangent of 0 ($startTan_{i=0} = 0$). The value for *startTan* and *startVal* for $i > 0$ depends on $t_{i-1,i}$, which is the time elapsed between FAP # bookmark$_{i-1}$ and FAP # bookmark$_i$. Thus, in accordance with one acceptable schema,

If $t_{i-1,i} > FAPdur_{i-1}$ then:

$startVal_i = FAPval_{i-1}$

$startTan_i = 0$

and the resulting amplitude of the FAP to be sent to the renderer is computed with equation (5):

$$FAPAmp_i(t) = startVal_i \cdot (2t^3 - 3t^2 + 1) + FAPval_i \cdot (-2t^3 + 3t^2) + startTan_i \cdot (t^3 - 2t^2 + 1) \quad (5)$$

with $t \in [0, 1]$

$FAPdur_i$ is used to relocate and scale the time parameter, t, from [0 1] to [t_i t_i+$FAPdur_i$] with t_i being the instant when the word following FAP # bookmark$_i$ in the text is pronounced. Equation (6) gives the exact rendering time:

Rendering time for $FAPAmp_i(t) = t_i + t \cdot FAPdur_i$. (6)

If $t_{i-1,i} < FAPdur_{i-1}$ then:

$startVal_i = FAPAmp_{i-1}(t_{i-1,i} / FAPdur_{i-1})$

$startTan_i = \tan_{i-1}(t_{i-1,i} / FAPdur_{i-1})$ which is computed with equation (3):

$$\tan_{i-1}(t) = startVal_{i-1} \cdot (6t^2 - 6t) + FAPval_{i-1} \cdot (-6t^2 + 6t) + startTan_{i-1} \cdot (3t^2 - 4t + 1) \quad (7)$$

with $t \in [0 \quad 1]$

and the resulting amplitude of the FAP is again computed with equation (5).

Thus, even if the user does not estimate properly the duration of each bookmark, the equation (4) function, more than any other function investigated, would yield the smoothest overall resulting curve.

The above disclosed a number of principles and presented an illustrative embodiment. It should be understood, however, that skilled artisans can make various modifications without departing from the spirit and scope of this invention. For example, while the functions described by equations (1) through (4) are monotonic, there is no reason why an expression from its beginning to its end must be monotonic. One can imagine, for example, that a person might start a smile, freeze it for a moment, and then proceed with a broad smile. Alternatively, one might conclude that a smile that is longer than a certain time will appear too stale, and would want the synthesized smile to reach a peak and then reduce somewhat. Any such modulation can be effected by employing other functions, or by dividing the duration into segments, and applying different functions, or different target magnitudes at the different segments.

Appendix

Table 1: FAP groups

Group	Number of FAPs
1: visemes and expressions	2
2: jaw, chin, inner lowerlip, cornerlips, midlip	16
3: eyeballs, pupils, eyelids	12
4: eyebrow	8
5: cheeks	4
6: tongue	5
7: head rotation	3
8: outer lip positions	10
9: nose	4
10: ears	4

Table 2: Visemes and related phonemes

Viseme #	phonemes	example	Viseme #	phonemes	example
0	none	na	8	n, l	lot, not
1	p, b, m	put, bed, mill	9	r	red
2	f, v	far, voice	10	A:	car
3	T,D	think, that	11	e	bed
4	t, d	tip, doll	12	I	tip
5	k, g	call, gas	13	Q	top
6	tS, dZ, S	chair, join, she	14	U	book
7	s, z	sir, zeal			

Table 3: Facial expressions defined for FAP 2.

#	expression name	textual description
1	joy	The eyebrows are relaxed. The mouth is open and the mouth corners pulled back toward the ears.
2	sadness	The inner eyebrows are bent upward. The eyes are slightly closed. The mouth is relaxed.
3	anger	The inner eyebrows are pulled downward and together. The eyes are wide open. The lips are pressed against each other or opened to expose the teeth.
4	fear	The eyebrows are raised and pulled together. The inner eyebrows are bent upward. The eyes are tense and alert.
5	disgust	The eyebrows and eyelids are relaxed. The upper lip is raised and curled, often asymmetrically.
6	surprise	The eyebrows are raised. The upper eyelids are wide open, the lower relaxed. The jaw is opened.

Table 4: FAP definitions, group assignments, and step sizes.

FAP names may contain letters with the following meaning: l = left, r = right, t = top, b = bottom, I = inner, o = outer, m = middle.

Column A is in units

Column B is in units or birectional

Column C is Positive Motion

Column D is FAP group, and

Columns E is Quantizer step size

#	FAP name	FAP description	A	B	C	D	E
1	viseme	Set of values determining the mixture of two visemes for this frame (e.g. pbm, fv, th)	na	na	na	1	1
2	expression	A set of values determining the mixture of two facial expression	na	na	na	1	1
3	open_jaw	Vertical jaw displacement (does not affect mouth opening)	MNS	U	down	2	4
4	lower_t_midlip	Vertical top middle inner lip displacement	MNS	B	down	2	2
5	raise_b_midlip	Vertical bottom middle inner lip displacement	MNS	B	up	2	2
6	stretch_l_cornerlip	Horizontal displacement of left inner lip corner	MW	B	left	2	2
7	stretch_r_cornerlip	Horizontal displacement of right inner lip corner	MW	B	right	2	2

8	lower_t_lip_lm	Vertical displacement of midpoint between left corner and middle of top inner lip	MNS	B	down	2	2
9	lower_t_lip_rm	Vertical displacement of midpoint between right corner and middle of top inner lip	MNS	B	down	2	2
10	raise_b_lip_lm	Vertical displacement of midpoint between left corner and middle of bottom inner lip	MNS	B	up	2	2
11	raise_b_lip_rm	Vertical displacement of midpoint between right corner and middle of bottom inner lip	MNS	B	up	2	2
12	raise_l_cornerlip	Vertical displacement of left inner lip corner	MNS	B	up	2	2
13	raise_r_cornerlip	Vertical displacement of right inner lip corner	MNS	B	up	2	2
14	thrust_jaw	Depth displacement of jaw	MNS	U	forward	2	1
15	shift_jaw	Side to side displacement of jaw	MNS	B	right	2	1
16	push_b_lip	Depth displacement of bottom middle lip	MNS	B	forward	2	1
17	push_t_lip	Depth displacement of top middle lip	MNS	B	forward	2	1
18	depress_chin	Upward and compressing	MNS	B	up	2	1

		movement of the chin (like in sadness)					
19	close_t_l_eyelid	Vertical displacement of top left eyelid	IRIS D	B	down	3	1
20	close_t_r_eyelid	Vertical displacement of top right eyelid	IRIS D	B	down	3	1
21	close_b_l_eyelid	Vertical displacement of bottom left eyelid	IRIS D	B	up	3	1
22	close_b_r_eyelid	Vertical displacement of bottom right eyelid	IRIS D	B	up	3	1
23	yaw_l_eyeball	Horizontal orientation of left eyeball	AU	B	left	3	128
24	yaw_r_eyeball	Horizontal orientation of right eyeball	AU	B	left	3	128
25	pitch_l_eyeball	Vertical orientation of left eyeball	AU	B	down	3	128
26	pitch_r_eyeball	Vertical orientation of right eyeball	AU	B	down	3	128
27	thrust_l_eyeball	Depth displacement of left eyeball	IRIS D	B	forward	3	1
28	thrust_r_eyeball	Depth displacement of right eyeball	IRIS D	B	forward	3	1
29	dilate_l_pupil	Dilation of left pupil	IRIS D	U	growing	3	1
30	dilate_r_pupil	Dilation of right pupil	IRIS D	U	growing	3	1
31	raise_l_i_eyebrow	Vertical displacement of left inner eyebrow	ENS	B	up	4	2

32	raise_r_i_eyebrow	Vertical displacement of right inner eyebrow	ENS	B	up	4	2
33	raise_l_m_eyebrow	Vertical displacement of left middle eyebrow	ENS	B	up	4	2
34	raise_r_m_eyebrow	Vertical displacement of right middle eyebrow	ENS	B	up	4	2
35	raise_l_o_eyebrow	Vertical displacement of left outer eyebrow	ENS	B	up	4	2
36	raise_r_o_eyebrow	Vertical displacement of right outer eyebrow	ENS	B	up	4	2
37	squeeze_l_eyebrow	Horizontal displacement of left eyebrow	ES	B	right	4	1
38	squeeze_r_eyebrow	Horizontal displacement of right eyebrow	ES	B	left	4	1
39	puff_l_cheek	Horizontal displacement of left cheeck	ES	B	left	5	2
40	puff_r_cheek	Horizontal displacement of right cheeck	ES	B	right	5	2
41	lift_l_cheek	Vertical displacement of left cheek	ENS	U	up	5	2
42	lift_r_cheek	Vertical displacement of right cheek	ENS	U	up	5	2
43	shift_tongue_tip	Horizontal displacement of tongue tip	MW	B	right	6	1
44	raise_tongue_tip	Vertical displacement of tongue tip	MW	B	up	6	1
45	thrust_tongue_tip	Depth displacement of tongue tip	MW	B	forward	6	1

46	raise_tongue	Vertical displacement of tongue	MW	B	up	6	1
47	tongue_roll	Rolling of the tongue into U shape	AU	U	concave upward	6	512
48	head_pitch	Head pitch angle from top of spine	AU	B	down	7	128
49	head_yaw	Head yaw angle from top of spine	AU	B	left	7	128
50	head_roll	Head roll angle from top of spine	AU	B	right	7	128
51	lower_t_midlip _o	Vertical top middle outer lip displacement	MNS	B	down	8	2
52	raise_b_midlip_o	Vertical bottom middle outer lip displacement	MNS	B	up	8	2
53	stretch_l_cornerlip_o	Horizontal displacement of left outer lip corner	MW	B	left	8	2
54	stretch_r_cornerlip_o	Horizontal displacement of right outer lip corner	MW	B	right	8	2
55	lower_t_lip_lm _o	Vertical displacement of midpoint between left corner and middle of top outer lip	MNS	B	down	8	2
56	lower_t_lip_rm _o	Vertical displacement of midpoint between right corner and middle of top outer lip	MNS	B	down	8	2

57	raise_b_lip_lm_o	Vertical displacement of midpoint between left corner and middle of bottom outer lip	MNS	B	up	8	2
58	raise_b_lip_rm_o	Vertical displacement of midpoint between right corner and middle of bottom outer lip	MNS	B	up	8	2
59	raise_l_cornerlip_o	Vertical displacement of left outer lip corner	MNS	B	up	8	2
60	raise_r_cornerlip _o	Vertical displacement of right outer lip corner	MNS	B	up	8	2
61	stretch_l_nose	Horizontal displacement of left side of nose	ENS	B	left	9	1
62	stretch_r_nose	Horizontal displacement of right side of nose	ENS	B	right	9	1
63	raise_nose	Vertical displacement of nose tip	ENS	B	up	9	1
64	bend_nose	Horizontal displacement of nose tip	ENS	B	right	9	1
65	raise_l_ear	Vertical displacement of left ear	ENS	B	up	10	1
66	raise_r_ear	Vertical displacement of right ear	ENS	B	up	10	1
67	pull_l_ear	Horizontal displacement of left ear	ENS	B	left	10	1
68	pull_r_ear	Horizontal displacement of right ear	ENS	B	right	10	1

We Claim:

1. A system comprising:

a decoder responsive to an input signal comprising text and FAP information, that separates the FAP information form the text, and develops phonemes from said text,

a converter responsive to said decoder, that converts said phonemes to additional FAP information and outputs said additional FAP information combined with said FAP information separated by said decoder, and

a face rendering module responsive to an applied face model signal and to said output developed by said converter.

Abstract

An enhanced arrangement for a talking head driven by text is achieved by sending FAP information to a rendering arrangement that allows the rendering arrangement to employ the received FAPs in synchronism with the speech that is synthesized. In accordance with one embodiment, FAPs that correspond to visemes which can be developed from phonemes that are generated by a TTS synthesizer in the rendering arrangement are not included in the sent FAPs, to allow the local generation of such FAPs. In a further enhancement, a process is included in the rendering arrangement for creating a smooth transition from one FAP specification to the next FAP specification. This transition can follow any selected function. In accordance with one embodiment, a separate FAP value is evaluated for each of the rendered video frames.

FIG. 1
PRIOR ART

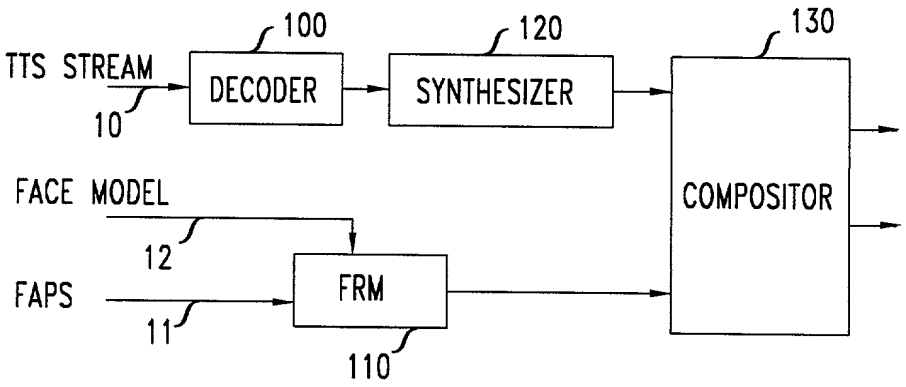


FIG. 2

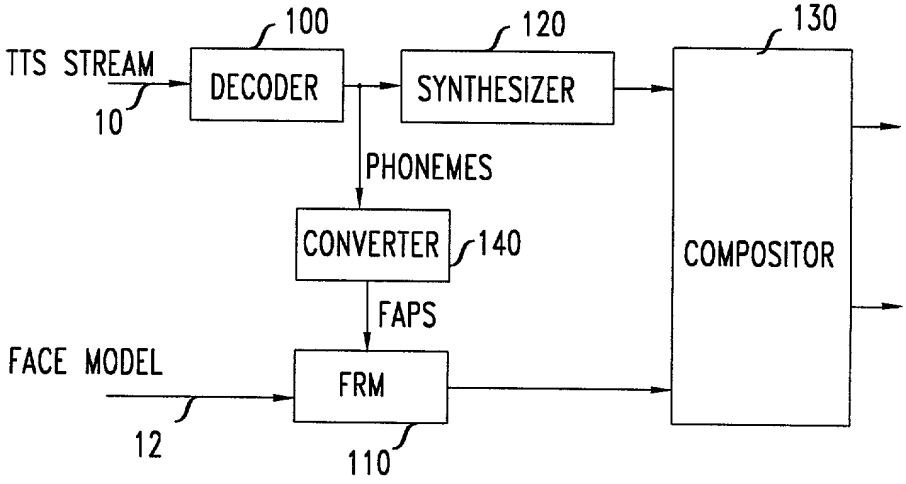
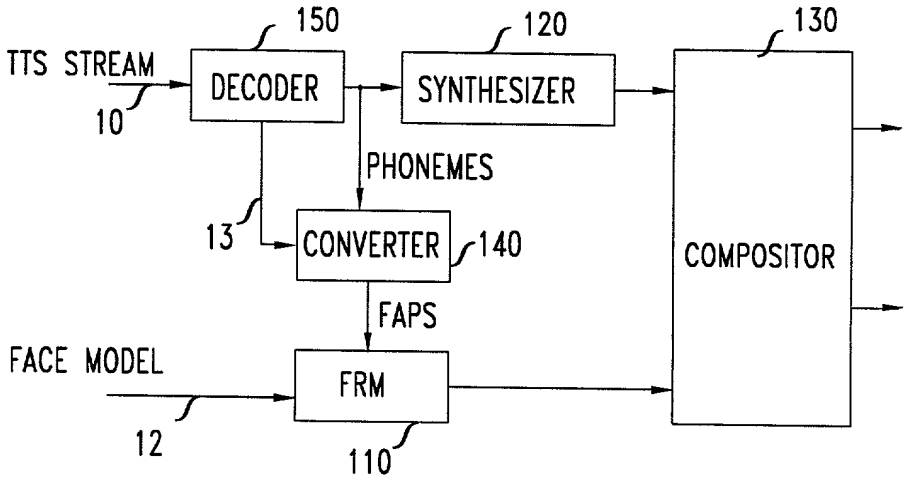


FIG. 3



IN THE UNITED STATES
PATENT AND TRADEMARK OFFICE

Declaration and Power of Attorney

As a below named inventor, I hereby declare that:

My residence, post office address and citizenship are as stated below next to my name.

I believe I am an original, first and sole inventor of the subject matter which is claimed and for which a patent is sought on the invention entitled **Integration of Talking Heads and Text-to-Speech Synthesizers for Visual TTS** the specification of which is attached hereto.

I hereby state that I have reviewed and understand the contents of the above identified specification, including the claims, as amended by an amendment, if any, specifically referred to in this oath or declaration.

I acknowledge the duty to disclose all information known to me which is material to patentability as defined in Title 37, Code of Federal Regulations, 1.56.

I hereby claim foreign priority benefits under Title 35, United States Code, 119 of any foreign application(s) for patent or inventors' certificate listed below and have also identified below any foreign application for patent or inventors' certificate having a filing date before that of the application on which priority is claimed:

None

I hereby claim the benefit under Title 35, United States Code, 120 of any United States application(s) listed below and, insofar as the subject matter of each of the claims of this application is not disclosed in the prior United States application in the manner provided by the first paragraph of Title 35, United States Code, 112, we acknowledge the duty to disclose all information known to us to be material to patentability as defined in Title 37, Code of Federal Regulations, 1.56 which became available between the filing date of the prior application and the national or PCT international filing date of this application:

Provisional application No. 60/082,393, filed April 20, 1998.

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

I hereby appoint the following attorney(s) with full power of substitution and revocation, to prosecute said application, to make alterations and amendments therein, to receive the patent, and to transact all business in the Patent and Trademark Office connected therewith:

Samuel H. Dworetsky	(Reg. No. 27873)
Thomas A. Restaino	(Reg. No. 33444)
Jose de la Rosa	(Reg. No. 34810)
Michele L. Conover	(Reg. No. 34962)
Robert B. Levy	(Reg. No. 28234)
Alfred G. Steinmetz	(Reg. No. 22971)
Benjamin S. Lee	(Reg. No. 42878)

I also appoint Henry T. Brendzel (Reg. No. 26,844) and William Ryan (Reg. No. 24,434) as associate attorneys, with full power to prosecute said application, to make alternations and amendments therein, and to transact all business in the Patent and Trademark Office connected therewith.

Please address all correspondence to Mr. S. H. Dworetsky, AT&T Corp., P.O. Box 4110, Middletown, New Jersey 07748. Telephone calls should be made to Henry T. Brendzel at (973) 467-2025.

Full name of joint inventor: Mark Charles Beutnagel

Inventor's signature _____ Date _____
Residence: Mendham, Morris County, NJ
Citizenship: USA
Post Office Address: 18 Mountain Avenue
Mendham, NJ 07945

Full name of joint inventor: Joern Ostermann

Inventor's signature _____ Date _____
Residence: Red Bank, Monmouth County, NJ
Citizenship: Germany
Post Office Address: 72 Walnut Avenue
Red Bank, NJ 07701

Full name of joint inventor: Schuyler Reynier Quackenbush

Inventor's signature _____ Date _____
Residence: Westfield, Union County, NJ
Citizenship: USA
Post Office Address: 744 Tamaques Way
Westfield, NJ 07090

Full name of joint inventor: Yao Wang

Inventor's signature _____ Date _____
Residence: Matawan, Monmouth County, NJ
Citizenship: China
Post Office Address: 69 Brandywine Drive
Matawan, NJ 07747